Composante du produit n° 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Estimation de la précision d'évolutions dans les enquêtes répétées, application à l'enquête suisse sur la valeur ajoutée

par Lionel Qualité et Yves Tillé

Décembre 2008

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, décembre 2008
Vol. 34, N° 2, pp. 193-201
Statistique Canada, N° 12-001-X au catalogue

Estimation de la précision d'évolutions dans les enquêtes répétées, application à l'enquête suisse sur la valeur ajoutée

Lionel Qualité et Yves Tillé [1]

Résumé

Nous proposons une méthode pour estimer la variance des estimateurs des évolutions qui prend en compte toutes les composantes de ceux-ci : le plan de sondage, le traitement des non-réponses, le traitement des grosses entreprises, la corrélation de la non-réponse d'une vague à l'autre, l'effet dû à l'utilisation d'un panel, la robustification et le calage au moyen d'un estimateur par le ratio. Cette méthode, qui permet la détermination d'intervalles de confiance des évolutions, est ensuite appliquée à l'enquête suisse sur la valeur ajoutée.

Mots clés : Covariance ; sondage stratifié ; panel.

1. Introduction

Dans les enquêtes longitudinales, la précision des évolutions dépend directement du taux de recouvrement des échantillons. Nous commençons par rappeler des résultats connus concernant les plans simples disjoints (voir à ce sujet Kish 1965 ; Sen 1973 ; Wolter 1985 ; Laniel 1988 ; Hidiroglou, Särndal et Binder (1995) ; Holmes et Skinner 2000 ; Nordberg 2000 ; Fuller et Rao 2001 ; Berger 2004). Ensuite nous calculons la variance des évolutions pour des plans simples dont les échantillons se superposent. Lorsque les taux de sondage sont très faibles la plupart de ces résultats sont bien connus et présentés par exemple dans Caron et Ravalet (2000). On peut trouver des résultats tenant compte des corrections de population finies dans Tam (1984).

Nous avons calculé précisément les variances des estimateurs pour une classe plus large de plans de sondage en population finie. Les corrections de population finie peuvent jouer un rôle important dans les enquêtes auprès des entreprises, car les entreprises de grandes tailles sont parfois sélectionnées avec des probabilités d'inclusion très élevées. Les calculs deviennent beaucoup plus compliqués en population finie pour la raison suivante : si la taille de la population est finie, deux échantillons disjoints ne sont pas indépendants. Si la population est infinie, deux échantillons indépendants sont disjoints. Plusieurs estimateurs sont examinés : la différence des estimateurs transversaux, la différence estimée uniquement sur la partie commune, les évolutions relatives. Les calculs deviennent encore plus complexes lorsque la population est dynamique (naissances, morts, changement de structure). La théorie que nous développons ci-dessous se limite au cas où la population ne change pas au cours du temps.

Dans la première partie, nous présentons le plan de sondage aléatoire simple bidimensionnel (voir à ce sujet Goga 2003) et nous donnons les estimateurs de Horvitz-Thompson correspondants. Nous calculons la variance de l'estimateur des évolutions basé sur ce plan de sondage. Dans une deuxième partie, nous donnons la variance d'autres estimateurs simples : l'évolution relative ou le quotient des totaux, l'estimateur de la différence basé sur l'intersection des échantillons. Nous décrivons ensuite comment ces résultats s'adaptent à la présence de non-réponse ignorable et à l'utilisation d'estimateurs plus complexes, qui font intervenir des poids modifiés pour obtenir des estimateurs calés, ou des variables modifiées par une procédure de robustification.

Ces résultats sur les plans simples sont facilement généralisables aux plans stratifiés à condition que les entreprises ne changent pas de strate d'une vague à l'autre. Enfin, nous appliquons cette méthode à l'enquête suisse sur la valeur ajoutée en prenant en compte toutes les composantes de l'enquête : la stratification, l'effet panel, la non-réponse, la corrélation entre les non-réponses d'une vague à l'autre, le calage au moyen d'un estimateur par ratio, et la robustification.

2. Estimation de la différence dans les plans simples

Soit une population $U = \{1, \ldots, k, \ldots, N\}$ de taille N dans laquelle sont prélevés deux échantillons s_1 et s_2 de tailles respectives n_1 et n_2. Ces échantillons peuvent avoir une partie commune (voir Figure 1).

On suppose que s_1 et s_2 sont des échantillons prélevés selon un plan simple sans remise, les tailles n_1 et n_2 ne sont donc pas aléatoires. Les échantillons s_1 et s_2 peuvent être décomposés en trois parties $s_A = s_1 \setminus s_2, s_B = s_2 \setminus s_1,$ et $s_C = s_1 \cap s_2$. Soit $n_A = |s_A|, n_B = |s_B|, n_C = |s_C|, n_1 = n_A + n_C, n_2 = n_B + n_C$. Les tailles de $s_A, s_B,$ et $s_C,$ peuvent être

1. Lionel Qualité et Yves Tillé, Institut de Statistique, Université de Neuchâtel, Pierre à Mazel 7, 2000 Neuchâtel, Suisse. Courriel : yves.tille@unine.ch.

aléatoires. Ce plan généralise, entre autres, les cas de figure suivants :

- si les échantillons s_1 et s_2 sont sélectionnés indépendamment, n_C est alors une variable aléatoire ;
- si l'échantillon s_1 est d'abord sélectionné, et l'échantillon s_2 est sélectionné dans le complémentaire de s_1 dans U alors s_C est vide et $n_C = 0$;
- si l'échantillon s_1 est d'abord sélectionné, et l'échantillon s_2 est constitué de l'union d'un sous-échantillon de taille fixe de s_1 et d'un échantillon de taille fixe du complémentaire de s_1 dans U, alors n_C n'est pas aléatoire, et l'on se retrouve dans le cas A de Tam (1984).

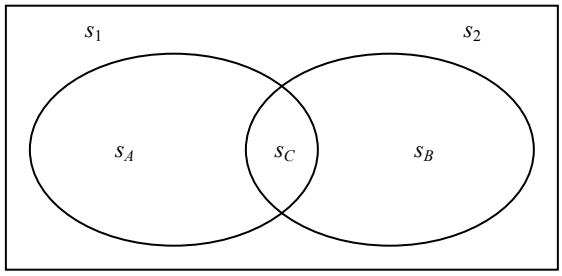

Figure 1 Échantillons qui se chevauchent

On fait l'hypothèse supplémentaire que conditionnellement à n_A, n_B, et n_C, les échantillons s_A, s_B, et s_C, sont simples, sans remise et de taille fixe. Ils proviennent du plan de sondage :

Définition 1. Plan simple bidimensionnel de taille fixe (n_A, n_B, n_C) :

$$p_{\text{simple}}(s_1, s_2 | n_A, n_B, n_C) = \begin{cases} \dfrac{n_A!\, n_B!\, n_C!\,(N - n_A - n_B - n_C)!}{N!} & \text{si } n_A = |s_A|, \\ & n_B = |s_B|, n_C = |s_C| \\ 0 & \text{sinon,} \end{cases}$$

où $s_A = s_1 \setminus s_2, s_B = s_2 \setminus s_1$ et $s_C = s_1 \cap s_2$ (voir à ce sujet Goga 2003).

La loi de tirage du couple (s_1, s_2), que l'on ne connaît pas en général, est donc supposée être de la forme

$$p(s_1, s_2) = p_{\text{simple}}(s_1, s_2 | n_A, n_B, n_C) \Pr(|s_1 \cap s_2| = n_C).$$

Soit deux variables x et y dont les valeurs prises sur les unités de U sont notées respectivement x_k et $y_k, k \in U$. Les variables x et y peuvent représenter la même variable mesurée à deux moments différents. On suppose également que x ne peut être observée que pour s_1 et y pour s_2. L'objectif est d'estimer les totaux

$$X = \sum_{k \in U} x_k \quad \text{et} \quad Y = \sum_{k \in U} y_k,$$

ainsi que la différence $Y - X$. Les estimateurs de Horvitz-Thompson de X et Y sont donnés par

$$\hat{X}_1 = \frac{N}{n_1} \sum_{k \in s_1} x_k \quad \text{et} \quad \hat{Y}_2 = \frac{N}{n_2} \sum_{k \in s_2} y_k.$$

2.1 Estimation naturelle de la différence

2.1.1 Variance de l'estimation de la différence

Une première manière de procéder pour estimer l'évolution $\Delta = Y - X$ est d'utiliser la différence des estimateurs transversaux $\hat{\Delta} = \hat{Y}_2 - \hat{X}_1$ qui est un estimateur sans biais conditionnellement à n_C sous le plan simple :

$$E(\hat{\Delta} | n_C) = Y - X,$$

et donc est également sans biais sous le plan p non-conditionnel à n_C.

Proposition 1 : La variance de $\hat{\Delta}$ vaut :

$$\begin{aligned} \operatorname{var}(\hat{\Delta}) = {} & N^2 \left(\frac{1}{n_1} - \frac{1}{N} \right) S_x^2 + N^2 \left(\frac{1}{n_2} - \frac{1}{N} \right) S_y^2 \\ & - 2N^2 \left(\frac{E(n_C)}{n_1 n_2} - \frac{1}{N} \right) S_{xy}, \end{aligned} \tag{1}$$

où

$$S_x^2 = \frac{1}{N-1} \sum_{k \in U} (x_k - \bar{X})^2, S_y^2 = \frac{1}{N-1} \sum_{k \in U} (y_k - \bar{Y})^2,$$

$$S_{xy} = \frac{1}{N-1} \sum_{k \in U} (x_k - \bar{X})(y_k - \bar{Y}).$$

La démonstration de (1) se trouve en annexe.

2.1.2 Cas particuliers et gain de précision

Le résultat (1) permet de traiter directement les cas particuliers de coordination suivants :

- si les deux échantillons forment un panel, $n_C = n_1 = n_2$, alors
$$\operatorname{var}(\hat{\Delta}) = N^2 \left(\frac{1}{n_C} - \frac{1}{N} \right) (S_x^2 + S_y^2 - 2S_{xy});$$
- si les échantillons sont disjoints (voir aussi Ardilly et Tillé 2003, pages 24-28) $n_C = 0$, et
$$\begin{aligned} \operatorname{var}(\hat{\Delta}) = {} & N^2 \left(\frac{1}{n_1} - \frac{1}{N} \right) S_x^2 + N^2 \left(\frac{1}{n_2} - \frac{1}{N} \right) S_y^2 \\ & + 2NS_{xy}. \end{aligned}$$
Il est surprenant de constater que la covariance ne dépend pas des tailles des échantillons. Elle est négative si x et y sont positivement corrélées, et devient négligeable par rapport aux termes de

variance quand la taille de la population est grande ;

– si q représente le taux de recouvrement fixé des deux échantillons et que $n_1 = n_2 = n$, on retrouve le cas A développé par Tam (1984). On obtient alors $n_C = qn$, et

$$\mathrm{var}(\hat{\Delta}) = N^2\left(\frac{1}{n}-\frac{1}{N}\right)(S_x^2+S_y^2)-2N^2\left(\frac{q}{n}-\frac{1}{N}\right)S_{xy};$$

– si les deux échantillons sont indépendants, $E(n_C) = n_1 n_2 / N$, et on retrouve

$$\mathrm{var}_{\mathrm{IND}}(\hat{\Delta}) = N^2\left(\frac{1}{n_1}-\frac{1}{N}\right)S_x^2 + N^2\left(\frac{1}{n_2}-\frac{1}{N}\right)S_y^2.$$

Si la taille de la population est grande et si les variables x et y ont des dispersions proches l'une de l'autre, le gain (ou la perte) de précision dû à la coordination par rapport à la sélection de deux échantillons de manière indépendante est

$$G = \frac{\mathrm{var}(\hat{\Delta})}{\mathrm{var}_{\mathrm{IND}}(\hat{\Delta})} \approx 1 - \rho q, \qquad (2)$$

où ρ est le coefficient de corrélation entre x et y, $\rho = S_{xy} / S_x S_y$ et q est le taux de recouvrement, $q = 2E(n_C)/(n_1 + n_2)$. L'expression (2) fournit un simple coefficient multiplicatif permettant de prendre en compte l'effet de la corrélation et du recouvrement.

2.1.3 Estimation de la variance de $\hat{\Delta}$

Pour estimer la variance, il faut considérer deux cas :

– si $E(n_C)$ est connu, ce qui peut être le cas (par exemple quand on sait que les deux échantillons sont indépendants), alors

$$\widehat{\mathrm{var}}(\hat{\Delta}) = N^2\left(\frac{1}{n_1}-\frac{1}{N}\right)s_{x1}^2 + N^2\left(\frac{1}{n_2}-\frac{1}{N}\right)s_{y2}^2 - 2N^2\left(\frac{E(n_C)}{n_1 n_2}-\frac{1}{N}\right)s_{xyC}, \qquad (3)$$

où

$$s_{x1}^2 = \frac{1}{n_1-1}\sum_{s_1}(x_k-\bar{x}_1)^2,\ s_{y2}^2 = \frac{1}{n_2-1}\sum_{s_2}(y_k-\bar{y}_2)^2,$$

et

$$s_{xyC} = \frac{1}{n_C-1}\sum_{s_C}(x_k-\bar{x}_C)(y_k-\bar{y}_C).$$

Cet estimateur est sans biais, mais il peut parfois prendre des valeurs negatives ;

– si $E(n_C)$ n'est pas connu, la seule information concernant la coordination est n_C.

$$\widehat{\mathrm{var}}(\hat{\Delta}) = N^2\left(\frac{1}{n_1}-\frac{1}{N}\right)s_{x1}^2 + N^2\left(\frac{1}{n_2}-\frac{1}{N}\right)s_{y2}^2 - 2N^2\left(\frac{n_C}{n_1 n_2}-\frac{1}{N}\right)s_{xyC}. \qquad (4)$$

Cet estimateur est sans biais conditionnellement à n_C et est donc aussi non-conditionnellement sans biais. Il peut aussi prendre parfois des valeurs négatives. Nous verrons plus loin que dans certaines applications où intervient de la non-réponse $E(n_C)$ n'est pas connu.

Pour utiliser l'estimateur (3), il est nécessaire d'avoir au moins deux unités dans l'intersection des échantillons $(n_C \geq 2)$, sauf si $E(n_C) = n_1 n_2 / N$. En effet, si $E(n_C) = n_1 n_2 / N$, ce qui est le cas quand les deux échantillons sont indépendants, le troisième terme de l'estimateur (3) est nul. L'estimateur (4) n'est quant à lui pas défini lorsque $n_C = 1$, sauf si $n_1 n_2 = N$.

2.2 Estimation au moyen de la partie commune

On peut aussi estimer la différence en utilisant uniquement la partie commune de l'échantillon, ce qui donne l'estimateur

$$\hat{\Delta}_C = N(\bar{y}_C - \bar{x}_C),$$

avec $\bar{y}_C = 1/n_C \sum_{k\in s_C} y_k$ et $\bar{x}_C = 1/n_C \sum_{k\in s_C} x_k$. Cet estimateur est sans biais non-conditionnellement et conditionnellement à n_C.

2.2.1 Estimation de la variance de $\hat{\Delta}_C$

La variance conditionnelle de $\hat{\Delta}_C$ vaut

$$\mathrm{var}(\hat{\Delta}_C | n_C) = N^2\left(\frac{1}{n_C}-\frac{1}{N}\right)(S_y^2 + S_x^2 - 2S_{xy}).$$

La variance non-conditionnelle vaut

$$\mathrm{var}(\hat{\Delta}_C) = N^2\left[E\left(\frac{1}{n_C}\right)-\frac{1}{N}\right](S_y^2 + S_x^2 - 2S_{xy}).$$

Cette variance non-conditionnelle peut être délicate à calculer quand n_C est aléatoire.

2.2.2 Comparaison des variances de $\hat{\Delta}$ et $\hat{\Delta}_C$

Si l'on veut comparer les deux estimateurs de la différence, on peut calculer

$$\mathrm{var}(\hat{\Delta}) - \mathrm{var}(\hat{\Delta}_C) = N^2\left[\frac{1}{n_1} - E\left(\frac{1}{n_C}\right)\right]S_y^2 + N^2\left[\frac{1}{n_2} - E\left(\frac{1}{n_C}\right)\right]S_x^2 - 2N^2\left[\frac{E(n_C)}{n_1 n_2} - E\left(\frac{1}{n_C}\right)\right]S_{xy}.$$

Si $n_1 = n_2 = n, S_x^2 = S_y^2 = S^2$, et $E(1/n_C) \approx 1/E(n_C)$, alors on obtient

$$\begin{aligned}\mathrm{var}(\hat{\Delta}) - \mathrm{var}(\hat{\Delta}_C) &\approx \frac{1}{qn}[q-1]\,2N^2S^2 - 2\frac{1}{qn}[q^2-1]\,\rho N^2 S^2 \\ &= \frac{2N^2S^2}{qn}(1-q)\,[\rho(1+q)-1],\end{aligned}$$

où $q = 2E(n_C)/(n_1+n_2)$ est le taux de recouvrement. L'estimateur $\hat{\Delta}_C$ est donc plus précis que $\hat{\Delta}$ si

$$\rho \geq \frac{1}{1+q}.$$

Par exemple, si $q = 0,7$, il est préférable de n'utiliser que la partie commune dès que $\rho \geq 1/(1+0,7) \approx 0,588$ (voir à ce sujet Caron et Ravalet 2000, page 346). Dans les cas où le recouvrement est important et la corrélation élevée, l'estimateur basé sur la différence des estimateurs transversaux n'est donc pas très pertinent.

3. Prise en compte de la non-réponse totale

On considère que la non-réponse est indépendante du plan de sélection. Selon le modèle, chaque unité décide de répondre ou non aléatoirement et les probabilités de réponse sont égales entre les unités. Il s'agit du modèle le plus élémentaire. Cependant, si une unité ne répond pas à la première vague, il est fortement probable qu'elle ne répondra pas non plus à la deuxième vague. Le modèle prend en compte cette dépendance en considérant séparément quatre cas :

- l'unité répond à la première vague et à la seconde ;
- l'unité répond à la première vague mais pas à la seconde ;
- l'unité ne répond pas à la première vague mais bien à la seconde ;
- l'unité ne répond ni à la première vague, ni à la seconde.

La non-réponse est couramment modélisée par un plan bernoullien multivarié, ce qui signifie que la probabilité de répondre est la même pour toutes les unités statistiques et également qu'une unité décide de répondre indépendamment de la réponse des autres unités. Le plan de non-réponse est le suivant :

$$q(r_A, r_B, r_C, r_D) = \phi_A^{\mathrm{card}\,r_A}\phi_B^{\mathrm{card}\,r_B}\phi_C^{\mathrm{card}\,r_C}\phi_D^{\mathrm{card}\,r_D},$$

où $r_A, r_B, r_C, r_D \subset U$, et r_A, r_B, r_C, r_D sont mutuellement exclusifs, et où

- $\phi_A^{\mathrm{card}\,r_A}$ est la probabilité de répondre à la vague 1, mais pas à la vague 2 ;
- $\phi_B^{\mathrm{card}\,r_B}$ est la probabilité de répondre à la vague 2, mais pas à la vague 1 ;
- $\phi_C^{\mathrm{card}\,r_C}$ est la probabilité de répondre à la vague 1, et à la vague 2 ;
- $\phi_D^{\mathrm{card}\,r_D}$ est la probabilité de ne répondre ni à la vague 1, ni à la vague 2.

La phase de non-réponse modélisée ainsi consiste donc en la sélection de quatre échantillons disjoints selon des plans bernoulliens avec des intensités différentes. Comme elle est supposée indépendante du plan de sondage, conditionnellement aux tailles d'échantillons observées, le plan résultant de la sélection et de la non-réponse est un plan simple multivarié. Si l'inférence est menée conditionnellement aux tailles d'échantillon, l'estimation des probabilités $\phi_A, \phi_B, \phi_C, \phi_D$ n'est pas nécessaire et une inférence sans biais peut être menée, comme si l'on avait affaire à un plan simple. La théorie de la section précédente s'applique donc directement sur les répondants, et toute l'information sur le recouvrement des deux échantillons se trouve dans $|s_C|$, que ce recouvrement soit dû au plan ou au lien existant entre les non-réponses aux deux vagues. Remarquons que même si le modèle est assez simple, il prend en compte le fait que si une unité n'a pas répondu à une vague, elle aura probablement moins de chance de répondre à la vague suivante. De plus, ce modèle sera appliqué dans des strates homogènes relativement petites.

4. Autres mesures des évolutions

La mesure de l'évolution n'est pas toujours exprimée en terme de différences. L'évolution est souvent mesurée sous forme de quotient, ou de différence relative. On considère donc les trois mesures suivantes :

- la différence $\hat{\Delta} = \hat{Y}_2 - \hat{X}_1$;
- l'évolution relative $\hat{\Delta}_R = (\hat{Y}_2 - \hat{X}_1)/\hat{X}_1 = \hat{Y}_2/\hat{X}_1 - 1$;
- le quotient $\hat{Q} = \hat{Y}_2/\hat{X}_1$.

La variance de $\hat{\Delta}$ peut s'exprimer simplement en fonction des estimateurs de variance de $\hat{Y}_2$ et $\hat{X}_1$ et de l'estimateur de leur covariance (voir expression 4). La variance de $\hat{\Delta}_R$ est égale à la variance de $\hat{Q}$. Elles peuvent être approchées puis estimées par une technique de résidus (voir à ce sujet Woodruff 1971 ; Binder et Patak 1994 ; Deville et Särndal 1992 ; Deville 1999),

$$\widehat{\mathrm{var}}(\hat{\Delta}_R) = \widehat{\mathrm{var}}(\hat{Q})$$
$$= \frac{1}{\hat{X}_1^2}\left[\widehat{\mathrm{var}}(\hat{Y}_2) + \hat{Q}^2\,\widehat{\mathrm{var}}(\hat{X}_1) - 2\hat{Q}\widehat{\mathrm{cov}}(\hat{X}_1, \hat{Y}_2)\right].$$

Cette variance peut donc être simplement estimée dès lors que l'on dispose d'estimateurs de $\mathrm{var}(\hat{Y}_2)$, $\mathrm{var}(\hat{X}_1)$ et $\mathrm{cov}(\hat{X}_1, \hat{Y}_2)$.

5. Estimation par ratio et robustification

Deux techniques sont couramment utilisées pour les estimations de résultats d'enquêtes par sondage : l'utilisation d'un estimateur par ratio pour caler sur le total d'une variable auxiliaire, et la robustification des estimateurs. Ces techniques doivent être prises en compte pour déterminer la précision des résultats finaux.

5.1 Calage

Si un estimateur est calé sur des totaux connus, la variance peut-être estimée simplement par une technique de résidus (voir Woodruff 1971 ; Binder et Patak 1994 ; Deville et Särndal 1992 ; Deville 1999). Par exemple, si $\mathbf{z}_{k1}$ et $\mathbf{z}_{k2}$ sont deux vecteurs colonnes de variables auxiliaires sur lesquelles les estimateurs $\hat{X}_{1\mathrm{Cal}}$ et $\hat{Y}_{2\mathrm{Cal}}$ sont calés en vagues 1 et 2, alors les variances peuvent être estimées par une technique de résidus : $\mathrm{var}(\hat{X}_{1\mathrm{Cal}}) \approx \mathrm{var}(\hat{E}_1)$ et $\mathrm{var}(\hat{Y}_{2\mathrm{Cal}}) \approx \mathrm{var}(\hat{E}_2)$, où $\hat{E}_1$ et $\hat{E}_2$ sont les estimateurs de Horvitz-Thompson des totaux des résidus, ces derniers étant donnés pour un plan simple et pour l'estimateur par la régression généralisée par :

$$e_{k1} = x_k - \mathbf{z}'_{k1}\hat{\mathbf{B}}_1,$$
$$e_{k2} = y_k - \mathbf{z}'_{k2}\hat{\mathbf{B}}_2,$$

avec

$$\hat{\mathbf{B}}_1 = \left(\sum_{k\in s_1} q_{k1}\mathbf{z}_{k1}\mathbf{z}'_{k1}\right)^{-1} \sum_{k\in s_1} q_{k1}\mathbf{z}_{k1}x_{k1},$$
$$\hat{\mathbf{B}}_2 = \left(\sum_{k\in s_2} q_{k2}\mathbf{z}_{k2}\mathbf{z}'_{k2}\right)^{-1} \sum_{k\in s_2} q_{k2}\mathbf{z}_{k2}y_{k2},$$

où $q_{kj}, j = 1, 2,$ est un coefficient qui permet de tenir compte d'une éventuelle hétéroscédasticité.

Dans le cas d'un plan de sondage à probabilités inégales, et par exemple d'un plan de sondage stratifié comme l'enquête suisse sur la valeur ajoutée, les résidus s'obtiennent en utilisant une régression pondérée. Il suffit de remplacer $\hat{\mathbf{B}}_1$ et $\hat{\mathbf{B}}_2$ respectivement par

$$\hat{\mathbf{B}}_1 = \left(\sum_{k\in s_1} \frac{q_{k1}\mathbf{z}_{k1}\mathbf{z}'_{k1}}{\pi_{k1}}\right)^{-1} \sum_{k\in s_1} \frac{q_{k1}\mathbf{z}_{k1}x_{k1}}{\pi_{k1}}, \quad \text{et} \qquad (5)$$

$$\hat{\mathbf{B}}_2 = \left(\sum_{k\in s_2} \frac{q_{k2}\mathbf{z}_{k2}\mathbf{z}'_{k2}}{\pi_{k2}}\right)^{-1} \sum_{k\in s_2} \frac{q_{k2}\mathbf{z}_{k2}y_{k2}}{\pi_{k2}}, \qquad (6)$$

où π_{kj} est la probabilité d'inclusion de l'unité k dans l'échantillon de la vague j, $j = 1, 2$.

5.2 Robustification

Il est souvent utile d'appliquer une technique de robustification qui permet de traiter les valeurs aberrantes. Considérons simplement que les valeurs aberrantes aient été détectées et que les poids des individus dont les valeurs sont considérées comme aberrantes aient été modifiés par un facteur $u_{kj}(s)$ à la vague j. Ce facteur est compris entre 0 et 1 et est égal à 1 pour les unités qui ont des valeurs considérées comme normales. La variance de l'estimateur robustifié peut être approchée en faisant l'hypothèse classique que les poids $u_{kj}(s)$ ne dépendent que faiblement de l'échantillon s qui a été tiré (voir Hulliger 1999). Il suffit alors de remplacer les variables x_k et y_k observées par $u_{k1}x_k$ et $u_{k2}y_k$ dans les estimateurs de variance.

En remettant ensemble toutes les composantes de l'erreur quadratique moyenne d'une évolution de manière à prendre en compte toutes les composantes de cette variance : le plan, l'effet panel, la non-réponse, le calage et la robustification, on obtient, pour l'évolution relative dans une strate,

$$\widehat{\mathrm{EQM}}(\hat{\Delta}_R) = \widehat{\mathrm{EQM}}(\hat{Q}) =$$
$$\frac{1}{\hat{X}_1}\left[\widehat{\mathrm{var}}(\widehat{EU}_1) + \hat{Q}^2\,\widehat{\mathrm{var}}(\widehat{EU}_1) - 2\hat{Q}\widehat{\mathrm{cov}}(\widehat{EU}_1, \widehat{EU}_2)\right], \quad (7)$$

où

$$\hat{X}_1 = \frac{N}{m_1}\sum_{R_1} x_k,\ \hat{Y}_2 = \frac{N}{m_2}\sum_{R_2} y_k,\ \hat{Q} = \frac{\hat{Y}_2}{\hat{X}_1},$$

$$eu_{k1} = u_{k1}x_k - u_{k1}\mathbf{z}'_{k1}\hat{\mathbf{B}}_1,$$
$$eu_{k2} = u_{k2}y_k - u_{k2}\mathbf{z}'_{k2}\hat{\mathbf{B}}_2,$$

$$\widehat{EU}_j = \frac{N}{m_j}\sum_{R_j} eu_{kj},\ \overline{EU}_j = \frac{\widehat{EU}_j}{N},\ j = 1, 2,$$

$$\hat{\mathbf{B}}_1 = \left(\sum_{k\in D_1} \frac{q_{k1}\,u_{k1}^2\,\mathbf{z}_{k1}\mathbf{z}'_{k1}}{\pi_{k1}}\right)^{-1} \sum_{k\in D_1} \frac{q_{k1}\,u_{k1}^2\,\mathbf{z}_{k1}\,x_k}{\pi_{k1}},$$

$$\hat{\mathbf{B}}_2 = \left(\sum_{k\in D_2} \frac{q_{k2}\,u_{k2}^2\,\mathbf{z}_{k2}\,\mathbf{z}'_{k2}}{\pi_{k2}}\right)^{-1} \sum_{k\in D_2} \frac{q_{k2}\,u_{k2}^2\,\mathbf{z}_{k2}\,y_k}{\pi_{k2}}.$$

$$\widehat{\mathrm{var}}(\widehat{EU_j}) = N^2\left(\frac{1}{m_j}-\frac{1}{N}\right)\frac{1}{m_j-1}\sum_{R_j}(eu_{kj}-\overline{EU}_j)^2,\ j=1,2,$$

$$\widehat{\mathrm{cov}}(\widehat{EU_1},\widehat{EU_2}) = N^2\left(\frac{m_C}{m_1m_2}-\frac{1}{N}\right)\frac{1}{m_C-1}\sum_{R_C}(eu_{k1}-\overline{EU_1}) \times(eu_{k2}-\overline{EU_2}).$$

R_1 et R_2 désignent l'ensemble des répondants à la première et à la deuxième vague dans la strate, $m_1=|R_1|$, $m_2=|R_2|$, $R_C=R_1\cap R_2$ et $m_C=|R_1\cap R_2|$. D_1 et D_2 sont les ensembles de répondants aux deux vagues dans le domaine dans lequel le calage a été réalisé.

6. L'enquête suisse sur la valeur ajoutée

6.1 Présentation de l'enquête

L'enquête suisse sur la valeur ajoutée est une enquête auprès des entreprises, réalisée chaque année. Elle vise à fournir des estimateurs des principaux paramètres de la production en Suisse : la valeur de production brute, le montant des consommations intermédiaires, la valeur ajoutée créée par les entreprises, et le coût de la main d'œuvre. Le plan de sondage utilisé est un échantillonnage stratifié d'entreprises. En 1999, un échantillon de 11 210 exploitations (occupant au moins deux personnes) a été sélectionné et sondé. Cet échantillon a été reconduit en 2000 et en 2001. Il s'agit donc sur cette période d'une enquête par panel. Faute d'un registre d'entreprises permettant d'identifier les naissances et les décès, la population des entreprises a été considérée comme constante pendant cette période. Le seul ajustement sur des données annuelles est réalisé au moyen d'une estimation par le ratio sur le total des Equivalents Temps Plein (ETP) par domaine d'activité, disponible par ailleurs.

La stratification est définie par les deux premiers chiffres de la NOmenclature Générale des Activités économiques (NOGA2) et par la taille de l'entreprise (voir Renfer 2000). Dans chaque strate d'activités, trois strates de taille sont constituées : les petites entreprises employant 2-19 personnes en ETP, les moyennes entreprises, de 20 à M ETP, et les grandes entreprises de plus de M ETP. La strate contenant les grandes entreprises est recensée, tandis que les petites et moyennes entreprises sont sélectionnées aléatoirement avec des taux de sondage différents. La borne M est choisie différemment dans chaque strate d'activités afin d'obtenir une précision optimale. À ces trois vagues, environ 6 000 établissements ont répondu. Le taux de réponse des grandes entreprises, qui devaient être recensées, était proche de 71 % et était plus élevé que celui des petites et moyennes entreprises. Il a été décidé a posteriori de traiter certaines très grosses entreprises séparément selon la méthodologie de la strate 'surprise' de Hidiroglou et Srinath (1981). On peut en effet penser que le taux de réponse a été meilleur pour les entreprises les plus grosses qui ont une structure administrative plus apte à répondre aux questions de l'enquête. Leur appliquer un poids égal à celui des autres grandes entreprises introduirait un biais ainsi qu'une variabilité trop importante. Les poststrates 'surprise' contiennent les 5 % d'exploitations les plus grandes dans le fichier d'enquête. Ces dernières ont alors été considérées comme effectivement recensées et ont reçu le poids 1. Aucun autre traitement (calage, robustification) ne leur sera appliqué. Les strates de tirage de petites, moyennes et grandes entreprises ont été mises à jour et certaines strates, (classes de taille) comptant peu d'exploitations ont été regroupées a posteriori. Si l'on accepte l'hypothèse que les très grandes entreprises ont effectivement été recensées, l'estimateur qui en découle est sans biais, et la variance liée aux très grandes entreprises est nulle. On peut donc calculer uniquement la variance dans les autres strates mises à jour.

Lors de l'enquête, la catégorie d'activités économiques était redemandée aux entreprises. Les estimations sont réalisées au niveau de ces NOGA2 déclarées et non au niveau des NOGA2 de la base de sondage. Un calage sur le nombre d'équivalent temps plein (ETP) donné par le registre des entreprises est ensuite réalisé au moyen d'un estimateur par le quotient au niveau des domaines NOGA2 'déclarés'.

Enfin, une technique de robustification a été utilisée pour écrêter la distribution de certaines variables dans l'échantillon des petites, moyennes et grandes entreprises (voir Hulliger 1999 ; Peters, Renfer et Hulliger 2001). Les poids des établissements dont les valeurs sont considérées comme aberrantes ont été modifiés par un facteur $u_{kj}(s)$ compris entre 0 et 1. Ce facteur est égal à 1 pour les entreprises qui ont des valeurs considérées comme normales.

6.2 Variance de l'évolution de la valeur ajoutée

L'objectif est d'estimer correctement la variance des estimateurs d'évolution de la valeur ajoutée (voir Renfer 2000 ; Peters *et al.* 2001). En calculant les variances sous une hypothèse d'indépendance des échantillons, on surestime largement les variances des évolutions parce que les variables « valeur ajoutée » aux temps t_1 et t_2 sont positivement corrélées. Une prise en compte correcte de tous les aspects du plan de sondage et du redressement devrait fournir de meilleures estimations de la précision. Les travaux portent sur les vagues d'enquête 1999, 2000 et 2001. Entre ces trois dates, l'échantillon brut n'a pas été modifié. Le fait que l'échantillon soit resté fixe devrait permettre d'estimer de manière fiable les évolutions, mais

un taux de réponse proche de 50 % peut faire perdre le bénéfice du panel, pour peu que le nombre de répondants communs aux vagues successives soit faible. Le cas de l'évolution entre deux vagues d'enquête où il y a eu une mise à jour de l'échantillon, et donc deux échantillons bruts et deux populations de référence différents constitue un tout autre problème.

Dans le cas présent, plusieurs raisons contribuent conjointement à l'obtention de faibles variances :

1. *Plan optimal* : Le plan de sondage a été optimisé. Selon la stratification optimale, les grosses entreprises ont des probabilités d'inclusion plus élevées. La strate d'entreprises contribuant le plus à la valeur ajoutée est recensée. Les estimateurs transversaux ont pour cette raison une faible variance.
2. *Fraction de réponse élevée* : Dans la strate recensée des grandes entreprises, le taux de réponse avoisine 70 %. La correction de population finie $(N-n)/N$ peut donc diviser la variance par 3 par rapport au cas d'une population infinie.
3. *Effet panel* : L'échantillon est un panel, ce qui est la meilleure stratégie pour estimer des évolutions.
4. *Corrélation de la non-réponse* : La non-réponse à une vague est très liée à la vague précédente, et ne dégrade donc pas beaucoup le panel.
5. *Corrélation des variables entre les vagues* : Les variables valeurs ajoutées au temps t et $t+1$ sont très corrélées, car il s'agit de la même variable mesurée à deux moments différents.
6. *Calage* : Les estimateurs sont calés dans les strates sur une variable liée à la variable d'intérêt, la variance des estimateurs peut alors s'écrire comme une variance résiduelle.

Sur les 11 210 entreprises sélectionnées en 1999, environ 5 200 ont répondu à la fois en 1999 et en 2000 ; et 5 300 ont répondu aux vagues 2000 et 2001. La taille du panel est donc relativement modeste, et le traitement de la non-réponse va avoir une grande influence sur les résultats. Afin de réaliser des estimations de variance, nous avons fait l'hypothèse que la non-réponse est ignorable (missing completely at random) au sein des strates de tirage.

À chaque vague, les estimations sont réalisées dans les domaines NOGA2 déclarés. Cela implique la possibilité d'un changement de domaine de la part des entreprises, qu'il faudrait essayer de prendre en compte dans les estimations longitudinales. Nous avons décidé de négliger l'impact de ces changements dans un premier temps, et de considérer pour l'estimation de la covariance que les domaines sont fixes et donnés par la valeur déclarée à la première des deux vagues consécutives. Cette simplification n'est pas abusive, dans la mesure où seules 30 entreprises (resp. 25) ont changé de domaine entre 1999 et 2000 (resp. 2000 et 2001), ce qui représente moins de 0,5 % (resp. 0,2 %) des ETP de l'échantillon. Le calage est réalisé pour chaque année, et peut être pris en compte au moyen d'une technique de résidus. Comme pour l'estimation de la précision des estimateurs transversaux, on prend en compte la robustification en repondérant les variables de l'enquête.

Moyennant des hypothèses réalistes, toutes les composantes de la variance peuvent être prises en compte grâce à l'expression générale (7). Cette expression est appliquée au sein de chaque strate et prend en compte toutes les composante de l'enquête sur la valeur ajoutée : l'effet panel, la non-réponse, la stratification, le calage, et la robustification. Les estimateurs de l'enquête sur la valeur ajoutée sont des estimateurs par ratio, et dans ce cas le calcul des résidus est simplifié. En effet, dans le cas du ratio, on calcule les coefficients de régression donnés en (5) et (6) en ayant une seule variable auxiliaire, donc $\mathbf{z}_{kj} = z_{kj}$ est scalaire. De plus, on prend $q_{kj} = 1/z_{kj}$, pour $j = 1, 2$, et on obtient alors en prenant en compte la robustification :

$$eu_{k1} = u_{k1}x_k - \hat{B}_1 u_{k1} z_{k1},$$
$$eu_{k2} = u_{k2}y_k - \hat{B}_2 u_{k2} z_{k2},$$

où

$$\hat{B}_1 = \frac{\sum_{D_1} u_{k1}x_k / \pi_{k1}}{\sum_{D_1} u_{k1}z_{k1} / \pi_{k1}},$$
$$\hat{B}_2 = \frac{\sum_{D_2} u_{k2}y_k / \pi_{k2}}{\sum_{D_2} u_{k2}z_{k2} / \pi_{k2}}.$$

6.3 Estimation de la précision des évolutions

Nous avons réalisé des estimations des écarts-types des évolutions des valeurs de productions brutes et des valeurs ajoutées calculées par l'Office fédéral suisse de la statistique. Ces estimations tiennent compte de tous les aspects développés précédemment. Nous les avons comparés aux écarts-types estimés qui auraient été obtenus sous l'hypothèse d'indépendance de tirage entre les vagues. Sur l'ensemble des strates d'activités, les écarts-types qui prennent en considération la corrélation entre les vagues d'enquête sont de 41 % inférieurs à ceux calculés sous hypothèse d'indépendance. Cela permet donc d'avoir des intervalles de confiance bien plus petits que ceux calculés avant ce travail, d'une manière plus rapide mais moins précise. Le gain n'est cependant pas le même dans toutes les strates d'activités. Dans les tableaux suivants sont donnés les écarts-types (ET) calculés pour les cinq plus grosses strates d'activités (NOGA) des évolutions de la valeur de

production brute (ΔVP) et de la valeur ajoutée (ΔVA) entre 1999 et 2000. L'écart-type qui aurait été obtenu en négligeant la corrélation entre les échantillons (ET_{ind}) est également inclus dans les tableaux, ainsi que le « gain » de précision réalisé en tenant compte de cette corrélation.

Tableau 1
Évolution de la valeur de production brute entre 1999 et 2000 et écarts-types (en milliards de francs suisses)

Strate	ΔVP	ET_{ind}	ET	Gain (en %)
1	3,31	2,35	0,87	63
2	-0,77	4,38	1,98	55
3	3,07	2,11	0,94	56
4	4,33	1,10	1,00	09
5	-0,09	0,81	0,53	35

Tableau 2
Évolution de la valeur ajoutée entre 1999 et 2000 et écarts-types (en milliards de francs suisses)

Strate	ΔVA	ET_{ind}	ET	Gain (en %)
1	1,96	0,91	0,32	65
2	0,68	2,99	1,04	65
3	1,90	1,47	0,72	51
4	0,36	0,47	0,45	05
5	-0,36	0,59	0,43	27

Remerciements

Ce travail a été réalisé dans le cadre d'une convention entre l'Université de Neuchâtel et l'Office fédéral suisse de la statistique. Les résultats publiés dans cet article n'engagent que les auteurs et en aucun cas l'Office fédéral de la statistique. Nous remercions Paul-André Salamin pour sa contribution à ce travail.

Annexe

Démonstration de la proposition 1

Il est bien connu que

$$\operatorname{var}(\hat{X}_1) = N^2\left(\frac{1}{n_1} - \frac{1}{N}\right)S_x^2$$

et

$$\operatorname{var}(\hat{Y}_2) = N^2\left(\frac{1}{n_2} - \frac{1}{N}\right)S_y^2.$$

Il suffit donc de calculer $\operatorname{cov}(\hat{X}_1, \hat{Y}_2)$. On note

$$\bar{x}_A = \frac{1}{n_A}\sum_{k\in s_A} x_k, \qquad \bar{x}_C = \frac{1}{n_C}\sum_{k\in s_C} x_k,$$

$$\bar{y}_B = \frac{1}{n_B}\sum_{k\in s_B} y_k, \qquad \bar{y}_C = \frac{1}{n_C}\sum_{k\in s_C} y_k,$$

$$\bar{x}_1 = \frac{n_A\bar{x}_A + n_C\bar{x}_C}{n_1}, \quad \bar{y}_2 = \frac{n_B\bar{y}_B + n_C\bar{y}_C}{n_2},$$

alors $\hat{X}_1 = N\,\bar{x}_1$ et $\hat{Y}_2 = N\,\bar{y}_2$. Il reste à calculer

$$\operatorname{cov}(\bar{x}_1, \bar{y}_2) = E\operatorname{cov}(\bar{x}_1, \bar{y}_2 \mid n_A, n_B, n_C) + \operatorname{cov}\left[E(\bar{x}_1 \mid n_A, n_B, n_C), E(\bar{y}_2 \mid n_A, n_B, n_C)\right].$$

Comme $\bar{x}_1$ et $\bar{y}_2$ sont sans biais conditionnellement à $n_A, n_B,$ et $n_C,$

$$\operatorname{cov}\left[E(\bar{x}_1|n_A, n_B, n_C), E(\bar{y}_2|n_A, n_B, n_C)\right] = \operatorname{cov}(\bar{X}, \bar{Y}) = 0.$$

On obtient donc

$$\operatorname{cov}(\bar{x}_1, \bar{y}_2) = E\operatorname{cov}(\bar{x}_1, \bar{y}_2|n_A, n_B, n_C).$$

Conditionnellement à $n_A, n_B,$ et $n_C,$ on est dans le cas A de Tam (1984, théorème 1). La variance conditionnelle vaut :

$$\operatorname{cov}(\bar{x}_1, \bar{y}_2|n_A, n_B, n_C) = \left(\frac{n_C}{n_1 n_2} - \frac{1}{N}\right)S_{xy}$$

et donc

$$\operatorname{cov}(\bar{x}_1, \bar{y}_2) = \left(\frac{E(n_C)}{n_1 n_2} - \frac{1}{N}\right)S_{xy}.$$

Or

$$\operatorname{cov}(\hat{X}_1, \hat{Y}_2) = N^2\operatorname{cov}(\bar{x}_1, \bar{y}_2),$$

ce qui permet d'obtenir le résultat (1).

Bibliographie

Ardilly, P., et Tillé, Y. (2003). *Exercices corrigés de méthodes de sondage*. Paris : Ellipses.

Berger, Y.G. (2004). Variance estimation for measures of change in probability sampling. *Canadian Journal of Statistics*, 32, 4, 451-467.

Binder, D.A., et Patak, Z. (1994). Use of estimating functions for estimation from complex surveys. *Journal of the American Statistical Association*, 89, 1035-1043.

Caron, N., et Ravalet, P. (2000). Estimation dans les enquêtes répétées : application à l'enquête Emploi en continu. Rapport technique 0005. Méthodologie Statistique, INSEE, Paris.

Deville, J.-C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes : linéarisation et techniques des résidus. *Techniques d'enquête*, 25, 219-230.

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Fuller, W.A., et Rao, J.N.K. (2001). Un estimateur composite de régression qui s'applique à l'Enquête sur la population active du Canada. *Techniques d'enquête*, 27, 49-56.

Goga, C. (2003). Estimation de la variance dans les sondages à plusieurs échantillons et prise en compte de l'information auxiliaire par des modèles nonparamétriques. Thèse de doctorat, Université de Rennes II, Haute Bretagne, France.

Hidiroglou, M., Särndal, C.-E. et Binder, D. (1995). Weighting and Estimation in Business Surveys. *Business Survey Methods*, (Éds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M. Colledge et P.S. Kott), New York : John Wiley & Sons, Inc., 477-502.

Hidiroglou, M.A., et Srinath, K.P. (1981). Some estimators of a population total from simple random samples containing large units. *Journal of the American Statistical Association*, 76, 690-695.

Holmes, D.J., et Skinner, C.J. (2000). Variance Estimation for Labour Force Survey Estimates of Level and Change. Technical report, Government Statistical Service Methodology Series, 21, Londres, Angleterre.

Hulliger, B. (1999). Simple and robust estimators for sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 54-63.

Kish, L. (1965). *Survey Sampling*. New York : John Wiley & Sons, Inc.

Laniel, N. 1988. (1988). Variances for a rotating sample from a changing population. *Proceedings of the Business and Economic Statistics Section*, American Statistical Association, 246-250.

Nordberg, L. (2000). On variance estimation for measure of change when samples are coordinated by the use of permanent random numbers. *Journal of Official Statistics*, 16, 363-378.

Peters, R., Renfer, J.-P. et Hulliger, B. (2001). Statistique de la valeur ajoutée : procédure d'extrapolation des données. Rapport technique, Office fédéral suisse de la statistique.

Renfer, J.-P. (2000). Enquête sur la production et la valeur ajoutée : échantillonnage complémentaire. Rapport technique, Office fédéral suisse de la statistique.

Sen, A.R. (1973). Theory and application of sampling on repeated occasions with several auxiliary variables. *Biometrics*, 29, 381-385.

Tam, S.M. (1984). On covariances from overlapping samples. *The American Statistician*, 38, (4), 288-289.

Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York : Spinger-Verlag.

Woodruff, R.S. (1971). A simple method for approximating de variance of a complicated estimate. *Journal of the American Statistical Association*, 66, 411-414.